Digital Humanities At Scale: Hathi Trust Research Center

Beth Plale

Co-Director, HathiTrust Research Center

Professor, School of Informatics and Computing

Indiana University

New questions require computational access to large corpus

Investigate way in which concepts of philosophy are used in physics through

- Extracting argumentative structure from large dataset using mixture of automated and social computing techniques
- Capture evidence for conjecture that availability of such analyses will enable innovative interdisciplinary research
- Digging into Data 2012 award, Colin Allen, IU

New questions, cont.

Document through software text analysis techniques, the appearance, frequency and context of terms, concepts and usages related to human rights in a selection of English-language novels.

- Ronnie Lipschutz of UCSC is currently doing this analysis on one of Jane Austen's books. He'd like to extend the work to encompass a far larger corpus.

New questions, cont.

Identify all 18th century published books in HathiTrust corpus, and apply topic modeling to create a consistent overall subject metadata

GOOGLE DIGITAL HUMANITIES AWARDS RECIPIENT INTERVIEWS REPORT
PREPARED FOR THE HATHITRUST RESEARCH CENTER

VIRGIL E. VARVEL JR.
ANDREA THOMER
CENTER FOR INFORMATICS RESEARCH IN SCIENCE AND SCHOLARSHIP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Fall 2011

The study

- Dr. John Unsworth, a representative of HTRC, distributed invitations to participate in this study via email to 22 researchers given Google Digital Humanities Research Awards.
- Interviews were conducted via telephone, Skype®, or face-to-face, and all were audio recorded. All participants agreed to IRB permission statement via email.
- A semi-structured interview protocol was developed with input from HTRC to elicit responses from participants on primary goals of project.

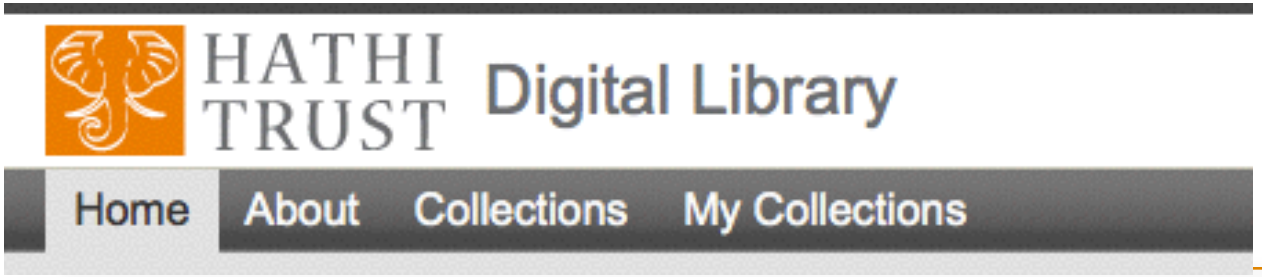
Select findings

- Optical Character Recognition
  - Steps should be taken to improve OCR quality if and when possible
  - Scalability of scanned image viewing is necessary for OCR reference and correction
  - Metadata should expose the quality of OCR

Select Findings

- “Would like better metadata about text languages, particularly in multi-text documents and on language by sections within text. Automatic language identification functions would be helpful, but human-created metadata is preferred, particularly for documents with low OCR quality.”
- “primary issue was retrieving the bibliographic records in usable form, unparsed by Google. […] process took 10 months to design the queries and get the data.”

HathiTrust Research Center: dedicated to provision of computational access to comprehensive body of published works for scholarship and education



Currently Digitized

10,100,278 total volumes
5,345,001 book titles
266,113 serial titles
3,535,097,300 pages
453 terabytes
120 miles
8,206 tons
2,784,331 volumes (~28% of total) in the public domain
View visualizations of HathiTrust
call numbers, languages, and dates

statistics information ››

→ HathiTrust is large corpus providing opportunity for new forms of computation investigation.
→ The bigger the data, the less able we are to move it to a researcher's desktop machine
→ Future research on large collections will require ***computation moves to the data, not vice versa***

Goal of HTRC

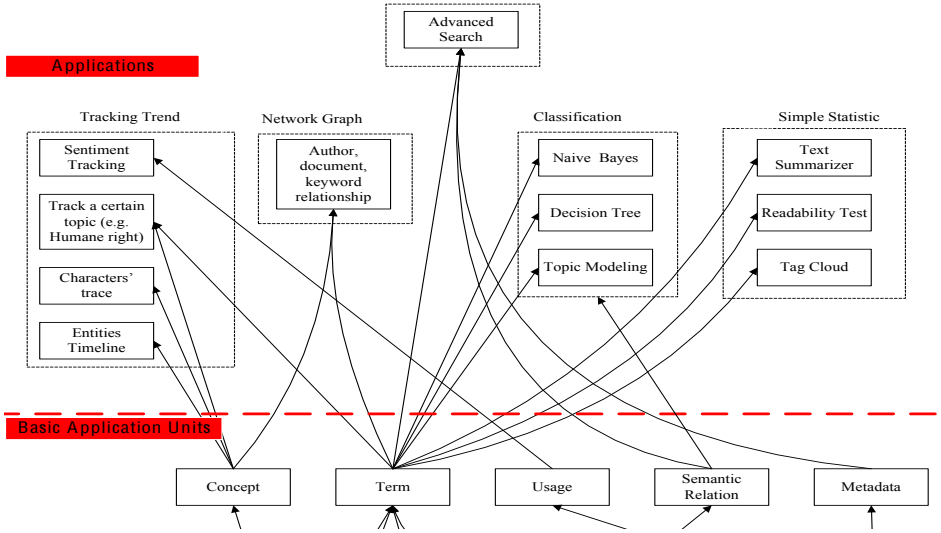
- HTRC will provide a persistent and sustainable structure to enable original and cutting edge research.
- Stimulate the development in community of new functionality and tools to enable new discoveries that would not be possible without the HTRC.

Goal, cont.

- Leverage data storage and computational infrastructure at Indiana U and UIUC,
- Provision secure computational and data environment for scholars to perform research using HathiTrust Digital Library.
- Center will break new ground, allowing scholars to fully utilize content of HathiTrust Library while preventing intellectual property misuse within confines of current U.S. copyright law.

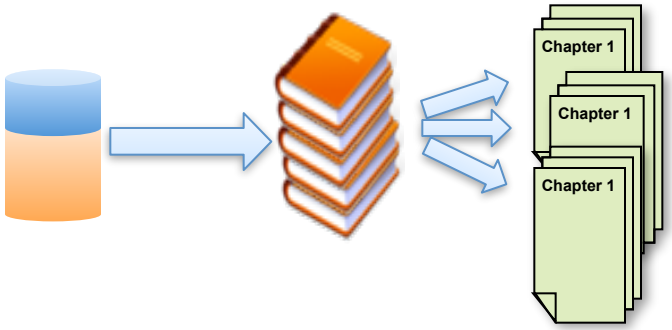
HathiTrust Research Center

- Analysis on 10,000,000+ volumes of HathiTrust digital repository
- Founded 2011
- Working with OCR
- Large-scale data storage and access
- HPC and Cloud

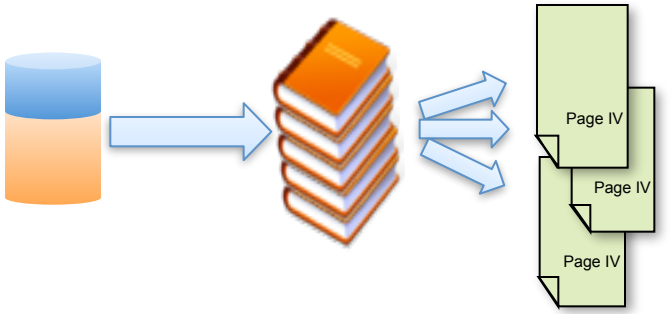


Type of Data (Public domain and copyrighted works)	Estimated initial size: 300-500 TB
Solr Indexes	36 TB (3 indexes)
File system rsync	12 TB
Fast volume access store	30TB
Versions of collection (5)	120 TB
Volume store indexes	100 TB

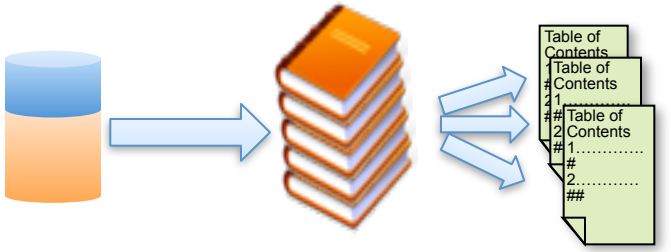
Corpus Usage Patterns



Access by chapter



Access by page



Access by special contents (table of contents, index, glossary)

HTRC Timeline

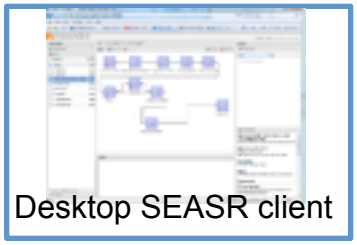
- Phase I: an 18-mo development cycle
  - Began 01 July 2011
  - Demo of capability June 2012 (12 mo mark)
- Phase II: broad availability of resource, begins 01 January 2013

Governance

- HTRC Exec Management Team: Beth Plale (IU), chair; Robert McDonald (IU), Marshall Scott Poole (UIUC), J. Stephen Downie (IU), John Unsworth (Brandeis Univ)
- Advisory board
- MOUs guide IU-UIUC interaction and HTRC-HT interaction
- Laine Farley, California Digital Library, and HT Executive Committee is liaison to HTRC
- Google Public Domain agreement – in process of signing (IU and UIUC individually executing)

What is it architecturally?

- Web services architecture and protocols
- Registry of services and algorithms
- Solr full text indexes
- noSQL store as volume store
- Large scale computing
- openID authentication
- Portal front-end, programmatic access
- SEASR mining algos



RESEARCH CENTER

Blacklight

CILogon (NCSA)

Agent framework

SEASR analytics service

WSO2 registry
services, collections, data capsule images

HTRC Data API v0.1

Access control (e.g. Grouper)

Agent
Agent
Agent instance
Agent instance

Solr index

Programmatic access e.g.,
PROJECT bamboo

Task deployment

Meandre Orchestration

Non-consumptive Data capsules

Volume store (Cassandra)

NCSA local resources

Future Grid

NCSA HPC resources

Penguin on Demand

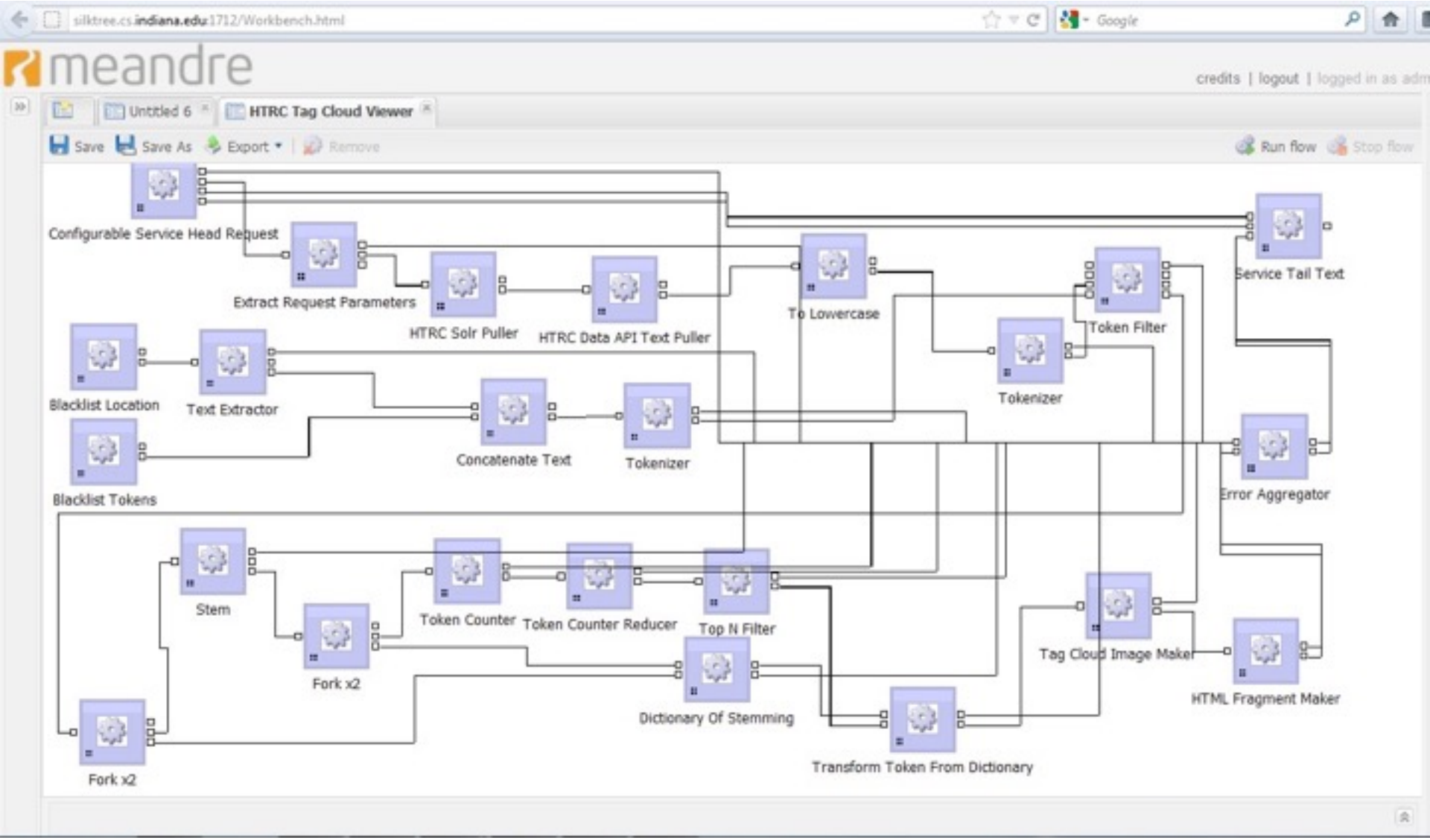
Page/volume tree (file system)

rsync

HathiTrust corpus
HATHI TRUST

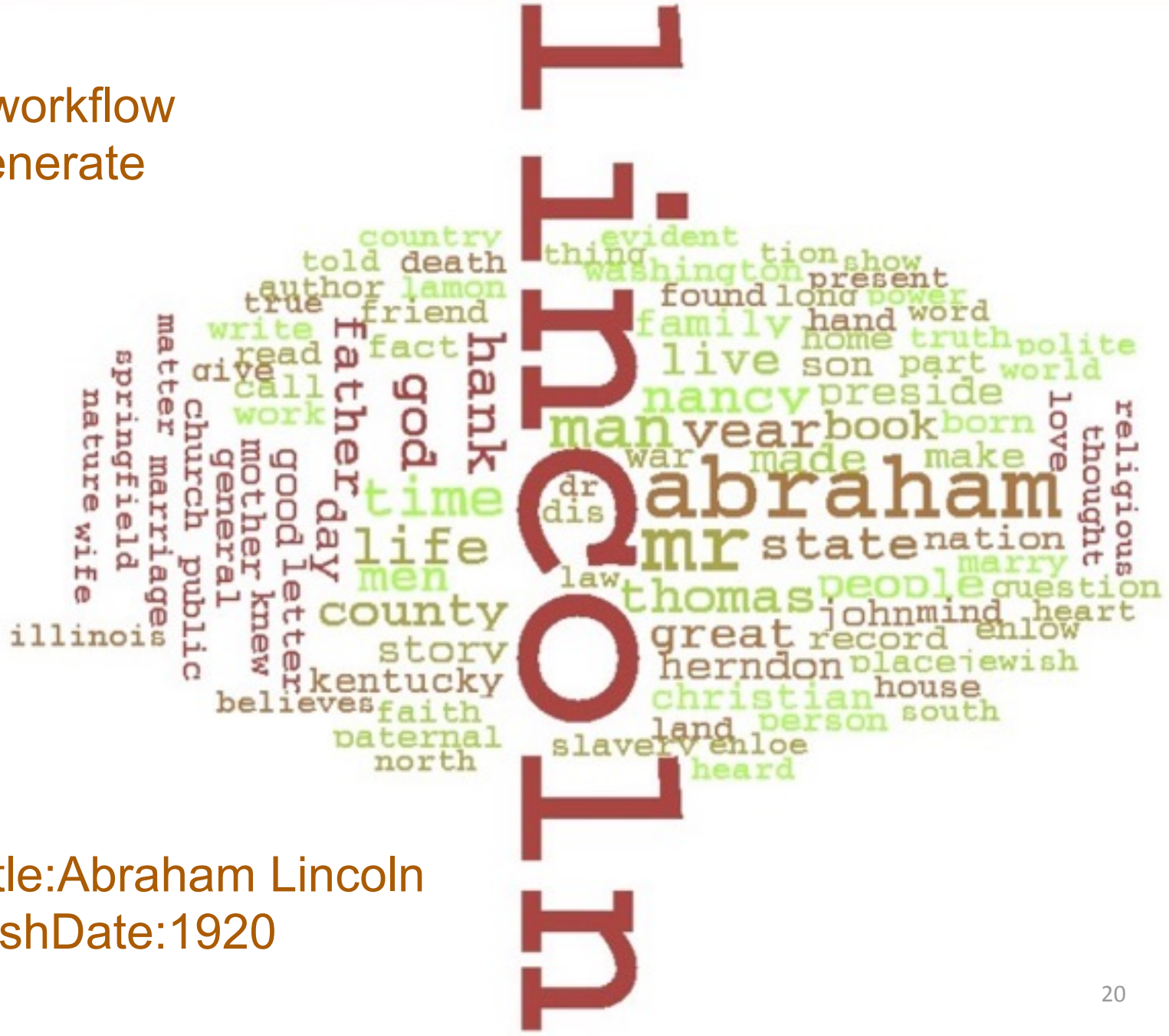
University of Michigan

One access point: through SEASR



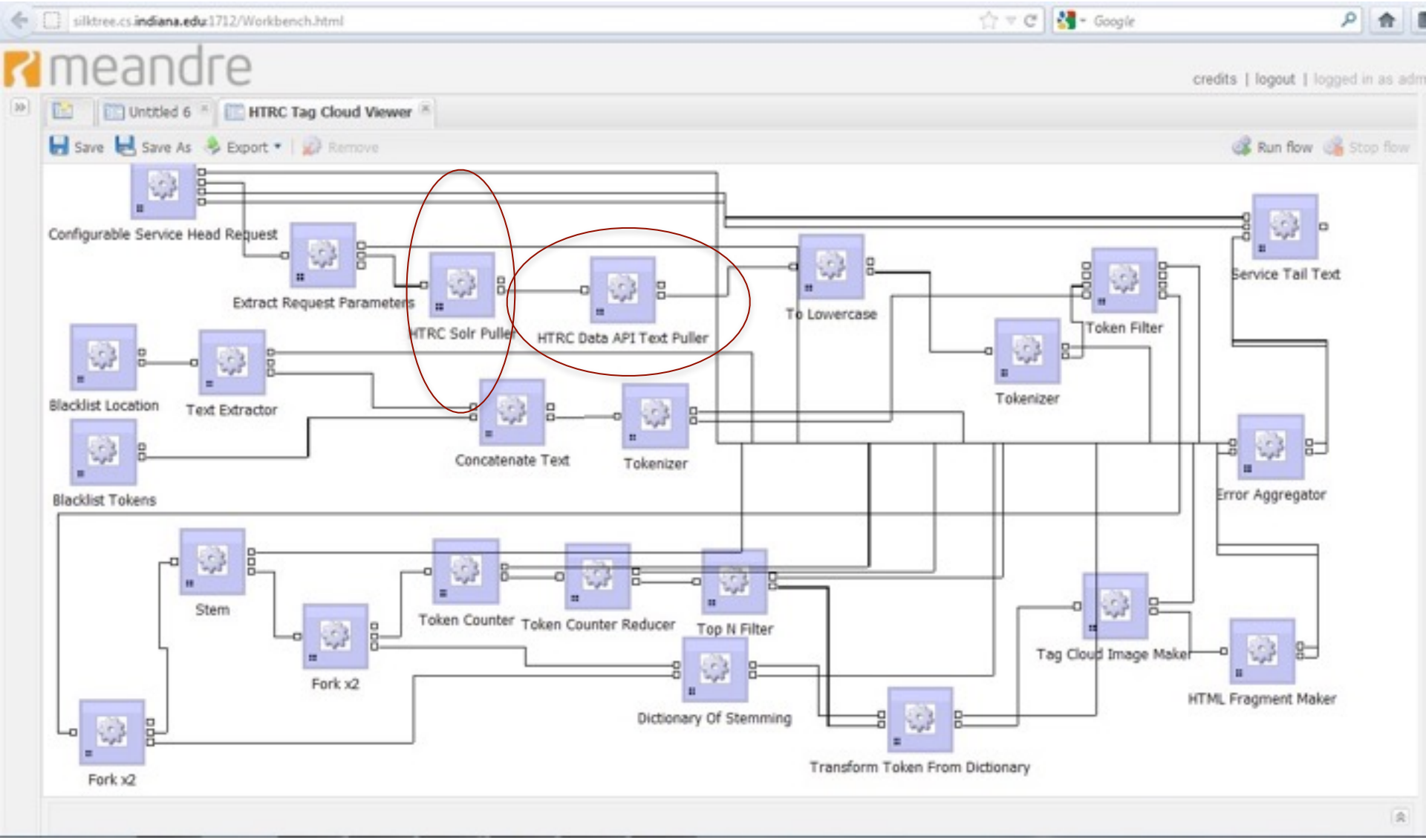
silktree.cs.indiana.edu:1716/service/data-api?query=title:Abraham Lincoln, AND publishDate:1920

SEASR: workflow used to generate tagcloud



Query = title:Abraham Lincoln AND publishDate:1920

Query = author:Withers, Hartley



Workflow invokes HTRC Solr index and HTRC data API.

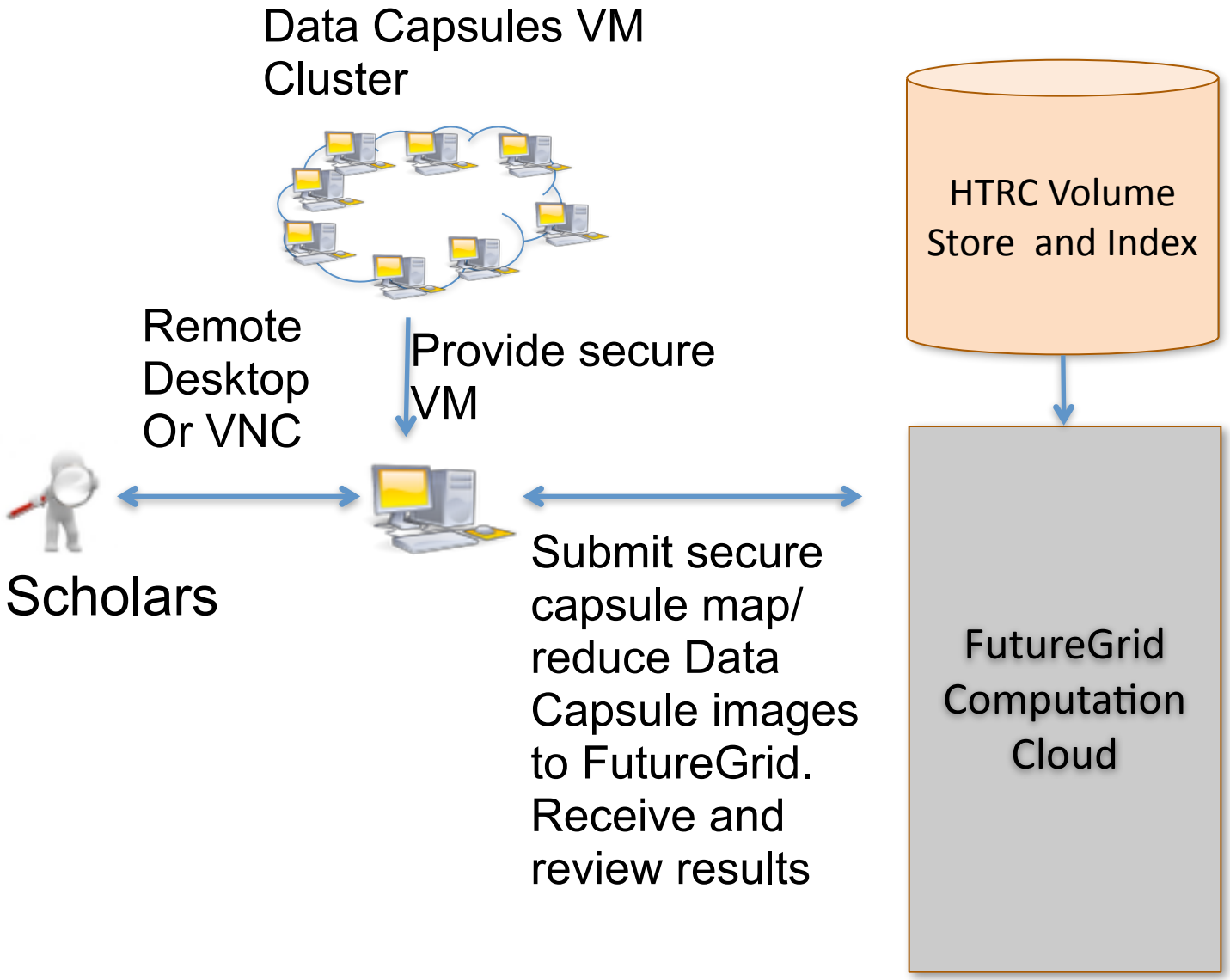
HTRC Solr index

- The Solr Data API 0.1 test version available.
  - Preserves all query syntax of original Solr,
  - Prevents user from modification,
  - Hides the host machine and port number HTRC Solr is actually running on,
  - Creates audit log of requests, and
  - Provides filtered term vector for words starting with user-specified letter
- Test version service soon available

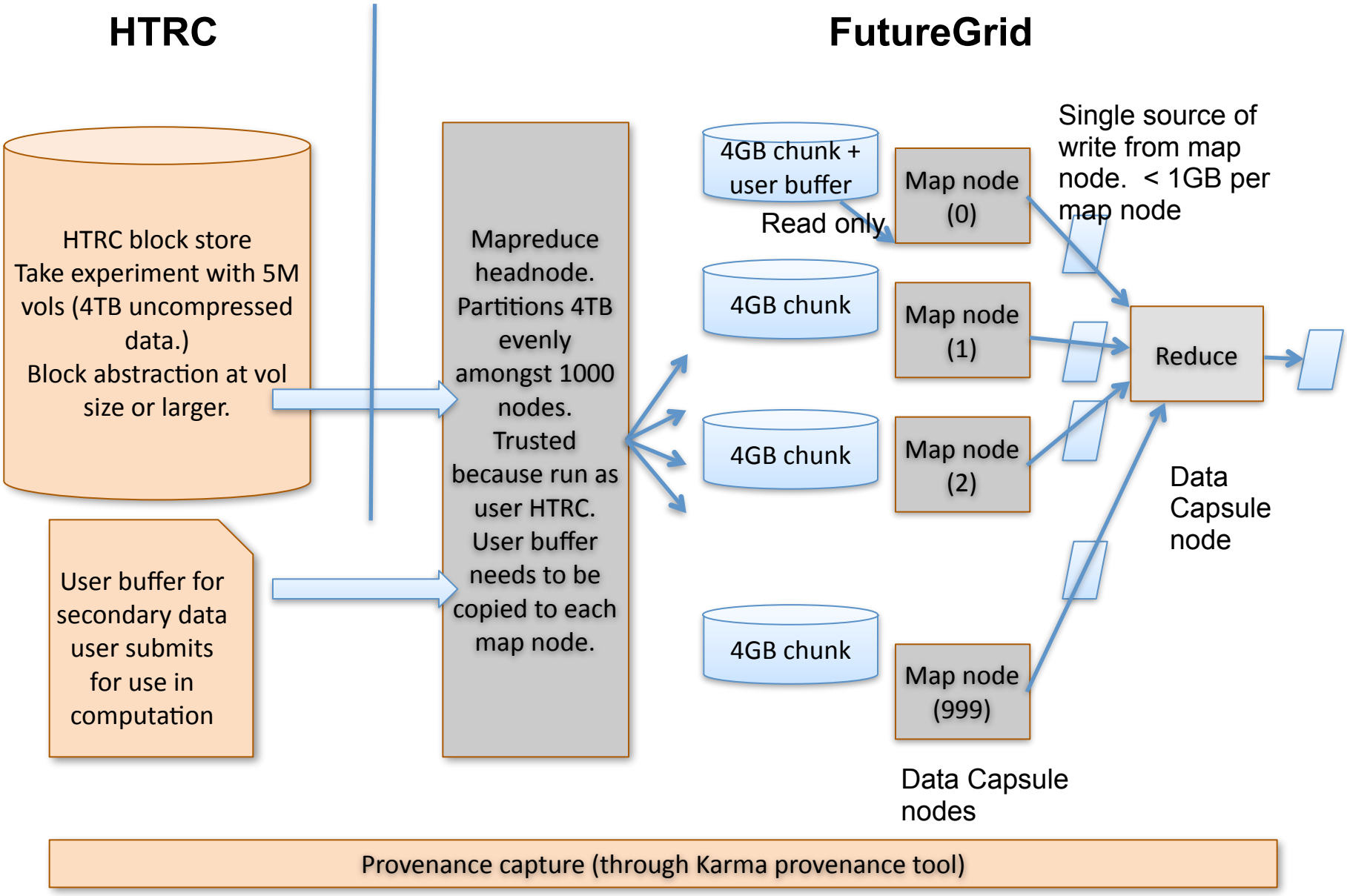
HTRC Solr index: most used API patterns

- ===== query
- http://coffeetree.cs.indiana.edu:9994/solr/select/?q=ocr:war
- =====faceted search
- http://coffeetree.cs.indiana.edu:9994/solr/select/?q=*:*&facet=on&facet.field=genre
- ===get frequency and offset of words starting with letter
- http://coffeetree.cs.indiana.edu:9994/solr/getfreqoffset/inu.32000011575976/w
- ===== banned modification request:
- http://localhost:8983/solr/update?stream.body=<delete><query>id:298253</query></delete>&commit=truethanks,

Requirement: run big jobs on large scale (free or nearly free) compute resources



Secure Data Capsule


HTRC
FutureGrid
HTRC block store
Take experiment with 5M vols (4TB uncompressed data.)
Block abstraction at vol size or larger.
User buffer for secondary data user submits for use in computation
Mapreduce headnode. Partitions 4TB evenly amongst 1000 nodes. Trusted because run as user HTRC. User buffer needs to be copied to each map node.
4GB chunk + user buffer
Read only
4GB chunk
4GB chunk
4GB chunk
Map node (0)
Map node (1)
Map node (2)
Map node (999)
Single source of write from map node. < 1GB per map node
Reduce
Data Capsule node
Data Capsule nodes
Provenance capture (through Karma provenance tool)

Purpose

FutureGrid allows researchers to experiment at all levels, including customizing network protocols and experimenting with new middleware. By using virtual machines running on real hardware, FutureGrid enables scientists to have full control over their testing environments without interfering with other users.

Scientists can also pursue interactive research and international collaboration, creating unprecedented potential for scientific discovery and innovation as they explore new uses for cloud technologies in science and engineering. Additionally, FutureGrid's availability to university students familiarizes the next generation of workers with these paradigms and their applications.

Acknowledgements

PI: Geoffrey Fox

CoPIs: Kate Keahey,Warren Smith, Jose Fortes, Andrew Grimshaw

Software Lead: Gregor von Laszewski

FutureGrid is sponsored by NSF grant #091081–FutureGrid:An Experimental, High-Performance Grid Test-bed. Partners in the FutureGrid project include: Indiana University, University of Chicago, Texas Applied Computing Center, University of Florida, University of Virginia, San Diego Supercomputer Center, Information Sciences Institute, Purdue University, University of Tennessee, and Technical University Dresden.

To learn how to join FutureGrid and apply for your own project, visit:

portal.futuregrid.org/gettingstarted

FutureGrid

An Experimental Grid Cloud and HPC Test-Bed

The FutureGrid project's mission is to enable experimental work that advances innovation and scientific understanding of distributed and parallel computing paradigms and:

- The engineering science of middleware that enables them
- Their use and drivers by important applications
- The education of new generations of students and personnel

About FutureGrid

FutureGrid is a distributed, high-performance test-bed that allows scientists to collaboratively create and test innovative approaches to parallel, grid, and cloud computing.

The test-bed comprises distributed clusters of high-performance computing resources connected to a high-speed network and linked to TeraGrid—the NSF's national cyberinfrastructure for scientific research.

The FutureGrid project is led by Indiana University and funded by the National Science Foundation.

Operational model

FutureGrid's operating model is different from both TeraGrid conventional clusters and commercial clouds, achieving flexibility by dynamically provisioning software onto "bare-metal" instead of loading images onto virtual machines. FutureGrid supports a growing Image Library that features platforms such as:

- MPI
- OpenMP
- Hadoop
- Dryad
- gLite
- Unicore
- Globus
- Xen

Typical FutureGrid projects

The flexibility in the configuration of FutureGrid resources enables its use across a variety of research and education projects. More than 120 projects are currently underway. The following are either complete or have attained significant achievements.

Project	Institution	Details
Educational Projects		
System Programming and Cloud Computing	Fresno State	Teaches system programming and cloud computing in different computing environments
REU: Cloud Computing	Arkansas	Offers hands-on experience with FutureGrid tools and technologies
Workshop: A Cloud View on Computing	Indiana School of Informatics and Computing (SOIC)	Boot camp on MapReduce for faculty and graduate students from underserved ADMI institutions
Topics on Systems: Distributed Systems	Indiana SOIC	Covers core computer science distributed system curricula (for 60 students)
Interoperability Projects		
SAGA	Louisiana State	Explores use of FutureGrid components for extensive portability and interoperability testing of Simple API for Grid Applications, and scale-up and scale-out experiments
Bio Application Projects		
Metagenomics Clustering	North Texas	Analyzes metagenomic data from samples collected from patients
Genome Assembly	Indiana SOIC	De novo assembly of genomes and metagenomes from next generation sequencing data
Non-Bio Application Projects		
Physics: Higgs boson	Virginia	Matrix Element calculations representing production and decay mechanisms for Higgs and background processes
Business Intelligence on MapReduce	Cal State - L.A.	Market basket and customer analysis designed to execute MapReduce on Hadoop platform
Computer Science Projects		
Data Transfer Throughput	Buffalo	End-to-end optimization of data transfer throughput over wide-area, high-speed networks
Elastic Computing	Colorado	Tools and technologies to create elastic computing environments using IaaS clouds that adjust to changes in demand automatically and transparently
The VIEW Project	Wayne State	Investigates Nimbus and Eucalyptus as cloud platforms for elastic workflow scheduling and resource provisioning
Technology Projects		
ScaleMP for Gene Assembly	Indiana Pervasive Technology Institute (PTI) and Biology	Investigates distributed shared memory over 16 nodes for SOAPdenovo assembly of Daphnia genomes
XSEDE	Virginia	Uses FutureGrid resources as a testbed for XSEDE software development
Globus Online	Indiana PTI, Chicago	Investigates the feasibility of providing DemoGrid and its Globus services on FutureGrid IaaS clouds

Creating Functionality around Non-consumptive Research

Key notions

Notion of a Proxy

- “researcher” in Google Book Settlement definition means a human
- An avatar is a virtual life form acting on behalf of person
- Computer program acts on behalf of person
- Computer program (i.e., proxy) must be able to read Google texts, otherwise computational analysis is impossible to carry out.
- So non-consumption applied to books applies to human consumption.

Non-consumptive research – implementation definition

- *No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.*
- Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.

Notion of Algorithm

- Computational analysis is accomplished through algorithms
  - An algorithm carries out one coherent analysis task: sort list of words, compute word frequency for text
- Researcher's computational analysis often requires running sequence of algorithms. Important distinction for implementing non-consumptive research is "who owns the algorithm"?

Infrastructure for computational analysis

- When needing to support computation over 10+M volume corpus, algorithms must be co-located with data.
- That is, algorithms must be located where repository is located, and not on user's desktop.
- When computational analysis is to be non-consumptive, likely one location for the data.

Who owns algorithm?

- HTRC owns the algorithms,
  - use Software Environment for Advancement of Scholarly Research (SEASR) suite of algorithms
  - we are examining security requirements of users, algorithms, and data

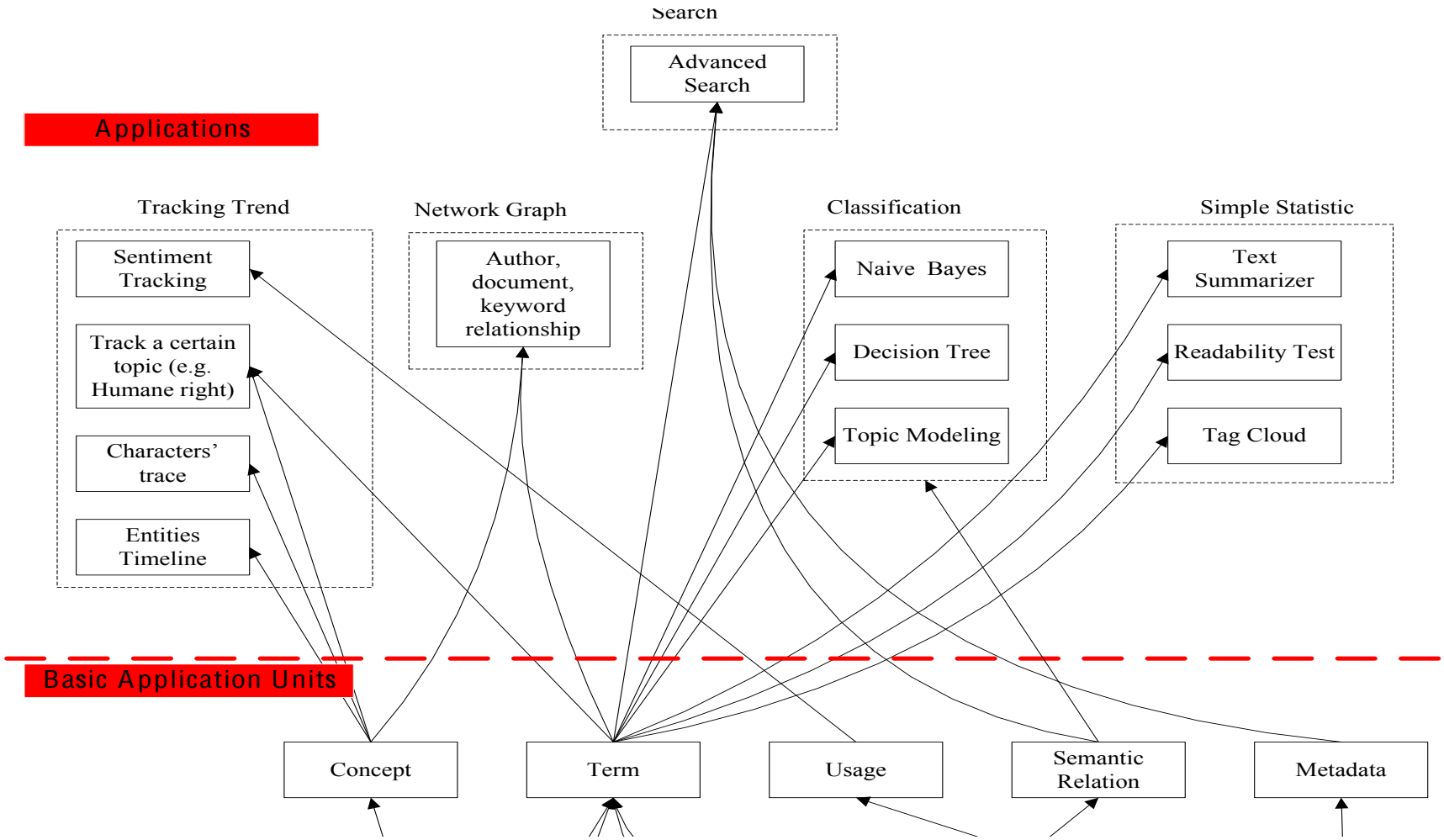
User owns and submits their algorithms

- HTRC recently received funding from Alfred P. Sloan foundation to prototype “data capsule framework” that provisions for non-consumptive research.
- Founded on principle of “trust but verify”. Informatics-savvy humanities scholar is given freedom to experiment with new algorithms on protected information, but technological mechanisms in place to prevent undesirable behavior (leakage.)

Non-consumptive, user-owned algorithms infrastructure; requirements:

- Implements non-consumptive
- Openness – users not limited to using known set of algorithms
- Efficiency – Not possible to analyze algorithms for conformance prior to running
- Low cost and scale – Run at large-scale and low cost to scholarly community of users
- Long term value –adoption for other purposes

Categories of algorithms. Can fair use be determined based on categorization of algorithm? Or is all computational use fair use?



Algo results fair use?

- Center supplied
  - Easier because we know category of algorithm
- User supplied
  - HTRC is not examining code, so open question

Parting philosophy

- Finally, results of computational research that conforms to restrictions of non-consumptive research must belong to researcher

How to Engage

- Building partnership with researchers and research communities is key goal of the HathiTrust Research Center
- HTRC can give technical advice to researchers as they look for funding opportunities involving access to research data
- Upcoming “Fix the OCR and Metadata Shortage Community Challenge” : help us address couple key weaknesses of HT corpus